Does the 'non-coding' strand code?

Volume 13 Number 4 1985 Nucleic Acids Research

Does the 'non-coding' strand code?

Paul M.Sharp

Department of Genetics, Trinity College, Dublin 2, Ireland

Received 4 December 1984; Revised and Accepted 28 January 1985

ABSTRACT

The hypothesis that DNA strands complementary to the coding strand contain in phase coding sequences has been investigated. Statistical analysis of the 50 genes of bacteriophage T7 shows no significant correlation between patterns of codon usage on the coding and non-coding strands. In Bacillus and yeast genes the correlation observed is not different from that expected with random synonymous codon usage, while a high correlation seen in 52 E.coli genes can be explained in terms of an excess of RNY codons. A deficiency of UUA, CUA and UCA codons (complementary to termination) seems to be restricted to the E.coli genes, and may be due to low abundance of the relevant cognate tRNA species. Thus the analysis shows that the non-coding strand has the properties expected of a sequence complementary to a coding strand, with no indications that it encodes, or may have encoded, proteins.

INTRODUCTION

Alff-Steinberger has suggested that there is "evidence for a coding pattern on the non-coding strand of the E.coli genome" (1). This conclusion was drawn from a compilation of codon usage figures for 52 E.coli genes, in which there exists "a significant positive correlation between the frequency with which a given codon appears on the coding strand and the frequency with which it appears, in phase, on the non-coding strand." This echoes an earlier report, from a separate group, that significant open reading frames exist in the same phase in the complementary DNA strands of genes from various sources (2). These 'coding properties' of the complementary strand, suggesting the existence of many heretofore cryptic genes, clearly warrant further investigation. If there is indeed widespread evidence of such a pattern it would have far-reaching implications for our understanding of the structure and evolution of genomes. One obvious possibility would be the involvement of these cryptic genes in regulation, while a correlation between codon usage in the two strands would necessarily constitute a (previously unidentified) constraint on synonymous codon choice.

Knowing that codon usage is non-random, and that patterns of codon usage

vary between species (3,4), two immediate questions arise: Firstly, is such a correlation, of coding patterns in the two strands, widespread in occurrence or peculiar to the particular data set investigated by Alff-Steinberger? Secondly, do the observed values for the correlation coefficient differ from those which might be expected, given that certain patterns of non-random codon usage have already been identified, and possibly explained?

To answer the first question codon usage data sets from bacteriophage T7, Bacillus spp. and the yeast Saccharomyces cerevisiae (nuclear genes only) are examined here. The T7 data set is particularly interesting, for several reasons. It comprises 50 genes and so is of a similar size to the E.coli data considered. Also T7 (a coliphage) depends on the host translational machinery (of E.coli) for replication. Finally, one reason for the existence of DNA sequences coding in both strands might be to encode many proteins within a DNA molecule of restricted length. Such a restriction is most likely to occur in viruses.

To answer the second question the expected value of the correlation coefficient must be determined. Alff-Steinberger has correctly stated that if codon usage were entirely random then a value of zero would be expected. However the requirement for genes to encode proteins with particular properties, and hence of particular amino acid composition, strongly suggests that codon usage can at best only be random within synonymous groups of codons (or possibly groups encoding biochemically similar amino acids (5,6)). The nature of the genetic code is such that, given non-random amino acid usage, random synonymous codon usage yields generally non-zero correlation coefficients (of codon usage in the two strands), with the exact magnitude dependent on the amino acid composition of the particular proteins encoded.

In fact, even synonymous codon usage has been found to be non-random in virtually all data sets so far examined. The much slower rate of evolution at the third position of codons, relative to that in pseudogenes is evidence that selective constraints on synonymous substitution exist (7,8). Several candidate constraints have been suggested (e.g. 9-15), though as yet no combination of these hypotheses has been found to be fully adequate (16,17). It is possible to calculate the expected correlation coefficient for codon usage under various constraints, in order to determine whether the observed values can be explained in terms of previously hypothesized selection pressures acting on codon usage. One explanation offerred by Alff-Steinberger (1) involves optimization of the codon-anticodon interaction energy (13), an apparent constraint on codon choice in E.coli (14) and bacteriophage T7 (16),

which might inflate the correlation due to use of energetically favoured triplet pairs. This constraint appears as third base pyrimidine bias - C is preferred to U when the initial doublet of the codon is A/U rich, while U is preferred to C if the doublet is C/G rich. Another property of coding sequences, perhaps the most ubiquitous so far reported, is a great excess of RNY (purine - any base - pyrimidine) codons (15). The manner in which patterns of codon preference might affect the correlation between the two strands can be discerned from examination of Table 1. Alff-Steinberger has wanted to examine the relationship between the usage figures for codons in panels (1),(2) - the mRNA - and those in panels (4),(3) - the same codons in the complementary strand. However, realizing that the figures for codons in panels (4) and (3) are fully determined by those for codons in panels (2) and (1) respectively she proceeded by calculating the correlation between panels (1) and (2). It can readily be seen that an excess of RNY codons in panels (1) and (2) would automatically lead to an excess of RNY codons in panels (3) and (4) (- the complement of an RNY codon is always also RNY). However third base pyrimidine bias in the mRNA leads to an excess in the complementary strand of a different subset of codons. It seems likely then that RNY bias (acting only in one strand), but not third base pyrimidine bias, could lead to an apparent correlation between the two strands. Note also that a correlation could arise in a random sequence if the G+C content were other than 0.5.

In this investigation the correlation coefficients expected under four models have been calculated for comparison with the observed values. The models are:

(1) Random synonymous codon usage.

(2) Random synonymous codon usage plus RNY bias.

(3) Random synonymous codon usage plus third base pyrimidine bias.

(4) Random synonymous codon usage plus RNY and third base pyrimidine bias.

Long open reading frames in complementary DNA strands would necessarily require avoidance, in coding sequences, of those triplets complementary to termination codons (i.e. UUA, UCA and CUA - see Table 1). The frequency of use of these three codons has also been investigated.

CODON USAGE DATA SETS

Codon usage data from E.coli was compiled by Alff-Steinberger from the EMBL Nucleotide Sequence Data Library, Release 2 (1). Data for bacteriophage T7 and Bacillus have been calculated previously (16,18), from references therein. The yeast data has been drawn from Release 4 of the EMBL Library

Table 1. Relationship of codon usage in two complementary strands.

mRNA (1)	mRNA (2)	complement (3)	complement (4)
UUU	UCU	AAA	AGA	AAA	AGA	UUU	UCU
UUC @	UCC	GAA	GGA	GAA @	GGA	UUC	UCC
UUA	UCA	UAA *	UGA *	UAA	UGA	UUA *	UCA *
UUG	UCG	CAA	CGA	CAA	CGA	UUG	UCG
CUU	CCU @	AAG	AGG	AAG	AGG @	CUU	CCU
CUC	CCC	GAG	GGG	GAG	GGG	CUC	CCC
CUA	CCA	UAG *	UGG	UAG	UGG	CUA *	CCA
CUG	CCG	CAG	CGG	CAG	CGG	CUG	CCG
AUU $	ACU $	AAU $	AGU $	AAU $	AGU $	AUU $	ACU $
AUC @$	ACC $	GAU $	GGU @$	GAU @$	GGU $	AUC $	ACC @$
AUA	ACA	UAU	UGU	UAU	UGU	AUA	ACA
AUG	ACG	CAU	CGU @	CAU	CGU	AUG	ACG @
GUU $	GCU @$	AAC @$	AGC $	AAC $	AGC @$	GUU @$	GCU $
GUC $	GCC $	GAC $	GGC $	GAC $	GGC $	GUC $	GCC $
GUA	GCA	UAC @	UGC	UAC	UGC	GUA @	GCA
GUG	GCG	CAC	CGC	CAC	CGC	GUG	GCG

Codons influenced directly ((1),(2)), and indirectly ((3),(4)) by bias in the mRNA strand:

$ - RNY bias (codons expected to be preferred).

@ - third base pyrimidine bias (codons expected to be preferred).

* - termination codons (expected to be rare).

(details available from the author on request). The size of the data sets is given in Table 2.

ANALYSIS

Following Alff-Steinberger the intraclass correlation of the frequency of use of codons in the coding and non-coding strands was calculated by pairing the usage figures for complementary codons (e.g. UUC and GAA) in one strand. The standard deviation of the correlation coefficient is taken to be 0.10 (Alff-Steinberger's bootstrap value (1)), although calculations using formulae presented in reference (19) suggest the somewhat higher value of 0.18. Expected correlation coefficients were calculated by first calculating expected codon usage data sets, as follows:

exp(1): Random synonymous codon usage was simulated by summing the observed codon usage within amino acids, and then assigning the average usage to each codon within the synonymous group. Necessarily the figures for UGG and AUG remain unchanged from the observed data set.

exp(2): Additional RNY bias was simulated by summing and averaging within

Table 2. Correlations of codon usage between coding and non-coding strands.

	n1	n2	P2	CORRELATION COEFFICIENT obs	exp(1)	exp(2)	exp(3)	exp(4)
E.coli	52	16351	0.62	0.448	0.135	0.325	0.095	0.284
T7	50	12145	0.62	0.075	0.080	0.284	0.059	0.229
Bacillus	17	5313	0.48	0.166	0.173	0.128	-	-
Yeast	19	7153	0.69	0.383	0.253	0.548	0.207	0.446

n1 - number of genes.
n2 - number of codons.
P2 - average third base pyrimidine bias (14).
exp(i) refers to model i (see text).

RNY and non-RNY codons (within synonymous groups) where appropriate.

exp(3): Additional third base pyrimidine bias was simulated by calculating average P2 values for each data set, and modifying the exp(1) values accordingly. (A P2 value is simply the proportional usage of codons of the favoured type, within the relevant pair of triplets (14)). Values were calculated for each pair of codons, and then averaged, to remove the effects of unequal amino acid usage. These P2 values are given in Table 2.

exp(4): RNY plus third base pyrimidine bias was simulated by modifying the exp(2) values by the appropriate P2 value.

For illustration the observed and expected usage figures, under each model, for the four Gly codons in the E.coli data set are presented in Table 3

RESULTS

Correlation coefficients are given in Table 2. The high value of 0.45 seen for the E.coli genes has already been reported (1). While the T7 data set is of a similar size to that for E.coli, the correlation coefficient for T7 is low, and not significantly different from zero. The values for the Bacillus and yeast data sets lie between the E.coli and T7 values.

Under the hypothesis that amino acid usage is non-random, but synonymous

Table 3. Example of expected codon usage figures (see text), in E.coli.

	obs	exp(1)	exp(2)	exp(3)	exp(4)
GGU	619	322.25	564.0	397.0	694.9
GGC	509	322.25	564.0	247.5	433.1
GGA	61	322.25	80.5	322.25	80.5
GGG	100	322.25	80.5	322.25	80.5

Table 4. Relative use of the synonymous codons for Leucine and Serine.

Leucine					Serine				
	E	T	B	Y		E	T	B	Y
UUA *	0.34	0.70	1.39	1.57	UCU	1.63	1.78	1.17	2.75
UUG	1.39	1.82	1.53	2.05	UCC	0.88	1.56	1.07	1.00
CUU	0.38	1.01	1.31	0.69	UCA *	0.49	0.83	1.15	0.75
CUC	0.38	0.65	0.40	0.19	UCG	0.80	0.36	0.68	0.30
CUA *	0.09	0.58	0.44	0.96	AGU	0.41	0.81	0.75	0.77
CUG	3.42	1.23	0.93	0.55	AGC	1.78	0.66	1.18	0.42

E: E.coli T: T7 B: Bacillus Y: Yeast
* codons complementary to termination.

codon usage is random, the expected value of the correlation coefficient varies widely, reflecting different patterns of amino acid usage in the data sets from different species. For both the T7 and Bacillus data sets the exp(1) value is slightly greater than the observed value.

In allowing for the known bias in use of RNY codons the value of exp(2) is greater than exp(1) in all but the Bacillus data set. The expected value for the yeast genes under this model is greater than that observed. While the observed value for E.coli is still greater than exp(2), the difference is no longer significant. Thirteen of these 52 E.coli genes have been noted to have high expression, and a more biassed pattern of codon usage (14). It is of interest to note that for these genes the observed correlation coefficient is lower (0.200).

The Bacillus genes (as a whole) show no evidence of third base pyrimidine bias (P2 values greater than 0.5 indicate the expected bias), and so it would be inappropriate to use models 3 and 4 for that data set. From Table 2 it can be seen that by incorporating third base pyrimidine bias into the models the correlation coefficient is always reduced (compare exp(3) and exp(4) with exp(1) and exp(2), respectively).

Usage figures for the triplets encoding Leu and Ser (which include the three codons that complement termination) are presented in Table 4. These are relative values calculated within species (1.00 indicating the expected usage), to aid comparison across species. The Leu codon CUA is consistently under-used, but only in the E.coli genes does there appear to be strong avoidance of the three codons as a group. Even then these 52 E.coli genes contain 203 'complementary stop' codons.

DISCUSSION

This consideration of codon usage data from four different sources

suggests that the correlation of coding and non-coding strands observed in E.coli (1) is not general. It is particularly striking that no evidence of a correlation is seen in a data set of a similar size from a coliphage.

It is also seen that a positive correlation, when observed, may be simply due to patterns of amino acid usage in the particular genes examined. The primary purpose of the genes considered is to encode a protein with particular properties which may depend critically on the precise amino acid sequence. While a small (and variable) proportion of nucleotide substitutions resulting in amino acid changes may be effectively neutral with respect to fitness, the majority will not (20). It is then naive to ignore amino acid usage and expect the correlation coefficient to be zero. For the T7 and Bacillus data sets the observed correlation is lower than that expected with random synonymous codon usage, and so there is no evidence there for a coding pattern on the non- coding strand.

In the E.coli data set the observed correlation is significantly higher than that for random synonymous codon usage. However, it is not significantly higher than that predicted by a model in which random synonymous codon usage is distorted by the RNY bias seen in the E.coli genes. The high correlation in yeast can also be explained by RNY bias. Two hypotheses (not mutually exclusive) have been proposed to explain this bias. Shepherd has observed an excess of RNY codons in protein coding genes from a wide variety of species (21), and has suggested that this may be a remnant of a primordial genetic code, in which only codons of the form RNY were used (15). Pieczenik, on the other hand, has considered the bases neighbouring the anticodon in the anticodon loops of E.coli tRNAs, and suggested that successive RNY codons would facilitate codon-anticodon interactions of more than three bases in length (22). He predicted that if such extended interactions are favourable then there should be an excess of RNY codons in the E.coli genome.

Although there is evidence of third base pyrimidine bias in the E.coli, T7 and yeast genes incorporating this bias into models of expected codon usage does not contribute to an explanation of the correlation coefficients observed. This was predicted from Table 1. However by taking account of just one previously identified constraint on synonymous codon choice, namely the widespread excess of RNY codons, the observed correlation coefficient in the E.coli genes can be explained. While it has not been ruled out that the excess of RNY codons in E.coli might result from codon preference in the complementary strand, the opposite cause and effect relationship seems far more likely. The existence of a strong bias towards use of RNY codons in T7

(16), in the absence of a high correlation coefficient, is just one piece of corroborative evidence. It is concluded then that there is no real evidence for a coding pattern in the non-coding strand of E.coli.

Avoidance of codons complementing termination does not appear to be widespread. Of the eight in phase "virtual genes" (open reading frames of greater than 100 codons, in the non-coding strand) previously reported (2) seven were from globin sequences, and six of these seven from mammalian sources. This may simply reflect non-use of the codons UUA, CUA and UCA in globin genes. Indeed this avoidance has been reported (23) and explained (in the case of UUA and UCA) as selection against codons liable to mutate to termination (23,24), although it seems more likely to be due to the relative abundance of corresponding tRNA species (25). In E.coli UUA and, particularly CUA are translated by Leu tRNAs of relatively minor abundance (12), and their low usage may be related to this. Open reading frames in the complementary strand, in phase, and of greater than expected length would then arise as an artefact. Again a previously hypothesized constraint on codon choice would appear to be sufficient to explain an 'unexpected' property of the complementary strand.

In conclusion, the suggestion that DNA strands presently considered to be non-coding (being complementary to identified genes) contain many sequences which encode, or may have encoded, cryptic genes is of great interest. However the hypothesis is not supported by this statistical analysis.

ACKNOWLEDGEMENT

I am most grateful to David J. McConnell for bringing reference (2) to my attention, for a great deal of stimulating discussion on codon usage in general, and for his comments on a draft of this paper in particular.

REFERENCES

1. Alff-Steinberger, C. (1984) Nucl. Acids Res. 12, 2235-2241.
2. Cascino, A., Cipollaro, M., Guerrini, A.M., Mastrocinque, G., Spena, A. and Scarlato, V. (1981) Nucl. Acids Res. 9, 1499-1518.
3. Grantham, R., Gautier, C., Gouy, M., Mercier, R. and Pave, A. (1980) Nucl. Acids Res. 8, r49-r62.
4. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. (1981) Nucl. Acids Res. 9, r43-r74.
5. Sneath, P.H.A. (1966) J. Theor. Biol. 12, 157-195.
6. French, S. and Robson, B. (1983) J. Mol. Evol. 19, 171-175.
7. Miyata, T. and Hayashida, H. (1981) Proc. Natl. Acad. Sci. USA 78, 5739-5743.
8. Li, W-H., Gojobori, T. and Nei, M. (1981) Nature 292, 237-239.
9. Clarke, B. (1970) Science 168, 1009-1011.
10. Richmond, R.C. (1970) Nature 225, 1025-1028.

11. Golding, G.B. and Strobeck, C. (1982) J. Mol. Evol. 18, 379-386.
12. Ikemura, T. and Ozeki, H. (1982) Cold Spring Harbor Symp. Quant. Biol. 47, 1087-1097.
13. Grosjean, H. and Fiers, W. (1982) Gene 18, 199-209.
14. Gouy, M. and Gautier, C. (1982) Nucl. Acids Res. 10, 7055-7074.
15. Shepherd, J.C.W. (1982) Cold Spring Harbor Symp. Quant. Biol. 47, 1099-1108.
16. Sharp, P.M., Rogers, M.S. and McConnell, D.J. (1985) J. Mol. Evol. (in press).
17. Warner, C. (1984) M.Sc. thesis, Trinity College, Dublin.
18. McConnell, D.J., Cantwell, B.A., Devine, K.D., Forage, A.J., Laoide, B.M., O'Kane, C., Ollington, J.F. and Sharp, P.M. (1985) Annals N.Y. Acad. Sci. (in press).
19. Snedecor, G.W. and Cochran, W.G. (1967) Statistical Methods, 6th edn. pp. 294-295, Iowa State University Press, Ames, Iowa.
20. Fitch, W.M. and Markowitz, E. (1970) Biochem. Genetics 4, 579-593.
21. Shepherd, J.C.W. (1981) J. Mol. Evol. 17, 94-102.
22. Pieczenik, G. (1980) Proc. Natl. Acad. Sci. USA 77, 3539-3543.
23. Modiano, G., Battistuzzi, G. and Motulsky, A.G. (1981) Proc. Natl. Acad. Sci. USA 78, 1110-1114.
24. Fitch, W.M. (1980) J. Mol. Evol. 16, 153-209.
25. Kimura, M. (1983) The Neutral Theory of Molecular Evolution. pp. 187-188, Cambridge University Press.